Predicting Fast Growing Firms

Fatima Arshad

2/26/2022

Introduction

Investment managers need to allocate funds to opportunities where the return can be maximized. The goal of this report is to classify companies into fast growing companies and companies not having a fast sales growth. This classification will eventually help the investment managers in well-informed decision making as to which companies to invest for maximum return.This report will discuss three models that predict probabilities of companies that would have a Compound Annual Growth Rate (CAGR) in sales of 40% or more between 2012 and 2014 and then classify the companies into two classes. This means that companies having a Compound Annual Growth Rate of 40% or more will be classified as a fast growing company. Input to these models are several features like income statement, balance sheet items which are necessary for an accurate prediction. The analysis will be done using the Logit, Logit LASSO, and Random Forest models with 5-fold cross validation. The accuracy and model selection of this analysis will be base upon the values of root mean squared error, the area under the curve, and the average expected loss.

Feature Engineering

The data used in this report is prepared by Bisnode and it has been sourced from the OSF website. The dataset is large containing observations of more than 287,000 rows with 48 explanatory variables in total. The time frame of the observations was from 2005 till 2016. For the classification of firms as fast growth we took a CAGR of 40% and calculated it based on 2012-2014 to account for stable growth. We identify and select firms which are currently operational, which is proven by the number of sales being greater than zero and not null. This analysis is only focused on small and medium sized companies i.e. whose sales are between the range of 1000 to 10 million euros. Furthermore, normalized values are created for all financial columns, income statement & balance sheet variables to ensure fair comparison across companies. Columns with flag variable were either imputed with mean value or replaced with zeros. After making these modifications, the analysis is based upon 10558 observations with 115 variables in total.

Our final choice of model building is the Random Forest(RF) Model used for building a stronger prediction model than logit counterparts. It is a black box model which is significantly good at classification, regression and other tasks. It operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. The predictors input to the random forest model are those input in Model 4 but without any feature engineering. Random forest returns the lowest 5-fold cross-validated RMSE of around 0.2957746 than all of the above mentioned models. It also returns the highest AUC of 0.7462442. Lowest RMSE and highest AUC suggest that prediction made by random forest model will be better than rest of the model given that we consider only these two parameters to assess the accuracy.

Loss Function

After gathering domain knowledge and carrying out research we define the loss function. The two important considerations made are the risk-free interest rate paid by depositing the money in a bank and the rate of

return on investing money in a company. The current interest rate provided by Hungarian banks on deposits as the risk-free rate is found to be3.3%. Next, we will carry out this analysis with the assumption that the rate of return on investment in a fast-growing company is 10%. This value is chosen because for stock market investments, anywhere from 7%-10% is usually considered a good ROI, and many investors use the S&P to guide their investment strategy. Furthermore, we create the loss function with the assumption that there will be 0% ROI if investment is made in a company that is non-fast-growing company.

Following the above methodology, we calculate the opportunity costs to arrive at the relative losses by false negatives and false positives. If the investment is made in a company and the classification was false positive, then the manager will lose the 3.3% return that could have been earned from depositing the money in a bank, hence the cost of a false positive is 3.3% risk free return.

On the contrary, if an investment is not made in the company based on a false negative classification, the loss would be (10% - 3.3%) 6.7% as the money is be deposited in a bank and the money will still earn 3.3% . Therefore, the ratio of cost of False Positive and False Negative turns out to be 1:2. It shows that false negative is twice as costly as the false positive cost.

Optimal Threshold & Classification

The optimal classification threshold based on these relative costs is 0.33. It is calculated using the optimal classification threshold formula which assumes that the model in use is the best one for prediction, which may not be true in practicality.

Therefore, we will calculate the optimal threshold using the data itself with incorporating our loss function. We plot the ROC curves to find the optimum threshold which turns out to be 0.35. Based on these classifications any company with a predicted probability of 0.35 or above will be classified as a fast-growing company.

Confusion Matrices

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It also tells the errors made by a classifier along with the type of error i.e. either false positive or false negative. We will examine the confusion matrix both, with and without the loss function. First, the confusion matrix without the loss function assigns the value of 1(interpreted as a fast-growing company) to any predicted probability of 0.5 or above. This value is not the optimum threshold as the losses from false positive and false negative are not always symmetric in the real world. Given that false negatives are more costly in our case i.e. the company is not fast-growing and still an investment is made, the goal is be to reduce the occurrence of false negatives in our predictions.

The *0.5 threshold* matrix has a percentage of false negatives is around 10.09% and percentage of false positives is 0.61%, whereas, with a *0.35 threshold* matrix, the percentage of false negatives is 9.37% and percentage of false positives is 2.7%. Based on our loss function, the model suggests that the company loses out around 1,176 Euros per firm and if the company evaluates 1000 firms in a year, the company loses out around 1.176 million Euros.

Conclusion

Based on th above prediction models, the random forest turns out to be the best model. Even though it is a black box model, it gives the best prediction accuracy, thus making it the optimum model for the investment management company. In order to further improve predictions and check for external validity, it is highly recommended to run these models for different time periods. Further more, we can also imporve our prediction by training our model on industry specific dataset rather than one large dataset. Additionally, having big data always improve the accuracy and machine learning, therefore, it is suggested to collect more observations on industry specific i.e. small and medium sized firms separately.

Link to RMD codes - GITHUB